

Hydrological Sciences Journal

ISSN: 0262-6667 (Print) 2150-3435 (Online) Journal homepage: http://www.tandfonline.com/loi/thsj20

Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe

Chantal Donnelly, Jafet C.M. Andersson & Berit Arheimer

To cite this article: Chantal Donnelly, Jafet C.M. Andersson & Berit Arheimer (2016) Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe, Hydrological Sciences Journal, 61:2, 255-273, DOI: 10.1080/02626667.2015.1027710

To link to this article: http://dx.doi.org/10.1080/02626667.2015.1027710

© 2015 The Author(s). Published by Taylor & Francis.

View supplementary material

Accepted author version posted online: 19 Mar 2015.
Published online: 14 Dec 2015.

Submit your article to this journal

Article views: 822

View related articles

View Crossmark data

Citing articles: 10 View citing articles

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=thsj20

Download by: [SMHI] **Date:** 22 February 2017, At: 01:44

HYDROLOGICAL SCIENCES JOURNAL – JOURNAL DES SCIENCES HYDROLOGIQUES, 2016
VOL. 61, NO. 2, 255–273
http://dx.doi.org/10.1080/02626667.2015.1027710

OPEN ACCESS

Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe

Chantal Donnelly, Jafet C.M. Andersson and Berit Arheimer

Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden

ABSTRACT

Open data make it possible to set up multi-basin models for large domains across environmental, climate and administrative boundaries. This study presents new methods for evaluating a number of aspects of multi-basin model performance, while exploring the performance of the E-HYPE_v2.1 model for several evaluation criteria in 181 independent river gauges across the European continent. Embedded model assumptions on dominant flow generating mechanisms are analysed by correlating physiographical characteristics to the flow regime. The results indicate that the model captures the spatial variability of flow and is therefore suitable for predictions in ungauged basins. The model shows good performance of long-term means and seasonality, while short-term daily variability is less well represented, especially for Mediterranean and mountainous areas. Major identified shortcomings refer to the resolution of precipitation patterns, aquifer exchanges, water extractions and regulation. This will guide the work with the next model version for which improvements in input data, processes and calibration have been identified to potentially contribute most to improved model performance.

ARTICLE HISTORY
Received 7 May 2014
Accepted 26 February 2015

EDITOR
D. Koutosyiannis

ASSOCIATE EDITOR
A. Efstratiadis

KEYWORDS
Pan-European modelling; open data; modified waters; ungauged basins; comparative hydrology; large-sample hydrology

1 Introduction

The magnitude and dynamics of river flow are the result of numerous interacting processes across many spatial and temporal scales within the river catchment (Blöschl *et al.* 2013, Hrachowitz *et al.* 2013). Catchments are therefore considered as complex systems (Rihani 2002, Kumar 2007) that have inherent limits to their predictability, e.g. due to unknown boundary conditions (e.g. Beven 2000). Predictions are even more uncertain in ungauged areas (e.g. Blöschl and Zehe 2005). Nevertheless, even if it is difficult to establish straightforward cause-and-effect relationships, catchments tend to generate emergent patterns in their hydrological response, which are useful for predictions (e.g. Wagener *et al.* 2007, 2013). Catchment characteristics reflect landscape features and the co-evolution of climate, geology, soils and vegetation (e.g. Berghuijs *et al.* 2014) as well as human alterations. Catchment characteristics determine the pathways and storage of water and, therefore, also the dominant processes and dynamics of river flow. Flow signatures can be quantified to summarize these river flow dynamics (Jothityangkoon *et al.* 2001, Eder *et al.* 2003). A reasonable assumption, which is tested in this study, is that catchments with similar physiographic characteristics generate similar flow signatures.

One way to learn more about the interactions between physiography and river flow, and the potential for predictions in ungauged basins, is to analyse many catchments in a synoptic and holistic way. Rather than studying one single catchment in detail, the idea is to compare many catchments with contrasting characteristics to understand the process controls on a complex system. This is referred to as *comparative hydrology* (Falkenmark and Chapman 1989, Sivapalan 2009, Blöschl *et al.* 2013) or *large-sample hydrology* (Gupta *et al.* 2014). In practice, this can be done by establishing a multi-basin model system to be used as a test-bench for hypotheses on process controls at many sites simultaneously (e.g. Arheimer *et al.* 2011).

A multi-basin hydrological model includes many river basins and gauges, crosses administrative boundaries, and a number of different geomorphological and climatic zones, often for very large domains (see for instance Liang *et al.* 1994, Alcamo *et al.* 2003, Thielen *et al.* 2009, Andersson *et al.* 2012, Strömqvist *et al.* 2012). Generally, the spatial distribution of topography, land cover and soil is used to simulate the spatially varying catchment behaviour over the model domain, using a

CONTACT Chantal Donnelly chantal.donnelly@smhi.se

Table S1 (Excel) for this article can be accessed at http://dx.doi.org/10.1080/02626667.2015.1027710

© 2015 The Author(s). Published by Taylor & Francis.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

distributed model. Given that the parameters of distributed models are linked to physiography, their optimization against a large sample of physiographic characteristics, in the context of the multi-basin modelling approach, becomes a type of regionalization. Predictions are thus also made for ungauged basins in the model domain because these basins are simulated using the same parameters for each physiographic class as the gauged basins (e.g. Donnelly *et al.* 2009, Arheimer and Lindström 2014). This is in contrast to currently popular regionalization approaches, in which gauged basins are used as donor catchments by statistically linking model parameters with catchment characteristics (e.g. Pechlivanidis *et al.* 2010, Carrillo *et al.* 2011).

In this paper we suggest that the assumption that there is a strong relationship between catchment characteristics and flow signatures should be tested in the model evaluation procedure, i.e. relationships in the observations should be assessed independently of the model, and then compared with the ability of the multi-basin model to reproduce these observed relationships. This exercise will indicate model shortcomings by indicating which processes may require more understanding. At the same time, an evaluation of the model performance at many sites across the domain, using a number of model performance metrics related to model purpose, will indicate where model shortcomings lie, further indicating the need for either better input data or process understanding implemented in model formulation or parameter improvements. The proposed test is summarized in the similarity concept of Fig. 1, which refers to (a) the similarity of observed *vs* modelled flow signatures and (b) the similarity of relationships between catchment characteristics and flow signatures, using observed *vs* modelled signatures.

We therefore assess whether a process-based model, calibrated with runoff generating parameters distributed according to soil type and land use, can provide useful information despite catchment uniqueness, i.e. we assume that our distributed parameter set is general for the entire domain. It has been argued that our ability to create a general hypothesis of hydrological behaviour is limited by catchment uniqueness (Beven 2000); however, the level of accuracy required for predictions in ungauged basins is also determined by the objective for which the model predictions are to be used. We explore this by widely testing the model against a dataset of independent discharge observations, to quantify the usefulness for predictions in ungauged basins. By doing this we elaborate and apply a new framework for model evaluation,

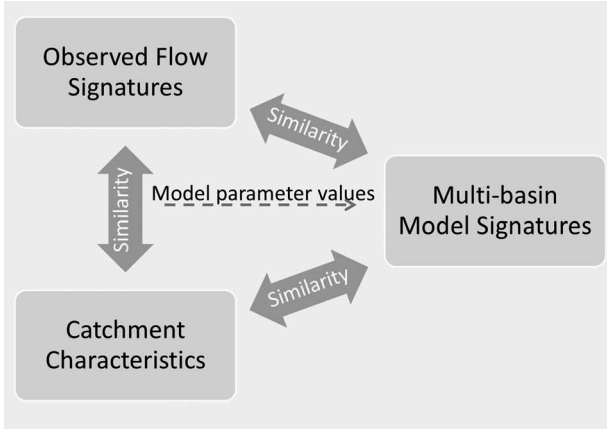


Figure 1. Similarities between catchment characteristics, observed flow signatures and multi-basin modelling.

which is based on the similarity concept of Fig. 1. Traditional evaluations of multi-basin hydrological models have often been limited to using few criteria and only discharge from large river basins (e.g. Döll *et al.* 2003, Hanasaki *et al.* 2008a, 2008b, Hagemann *et al.* 2009, Thielen *et al.* 2009). This new framework takes into account the multiple uses of large-scale model data and therefore includes multiple criteria, evaluated across multiple spatial and temporal scales and several model aspects, such as basic model assumptions on catchment similarities, flow signatures, different parts of the hydrograph, and both spatial and temporal variability.

In this study, the direct similarity between flow signatures and catchment characteristics is quantified for 181 non-pristine basins, ranging from 5000 to 800 000 km^2 across Europe. The catchments cover a wide variety of climatological and physiographical regions, with anthropogenic impacts. In the paper, we pose the following scientific questions:

(1) How strong and significant are the correlations between flow signatures and catchment physiography for a wide range of rivers across Europe?
(2) To what extent can homogenous, multi-basin models based on catchment characteristics help to define flow signatures for ungauged rivers across Europe?
(3) How can the similarity concept be applied to evaluate process-based models?

These questions are addressed by determining (a) which catchment characteristics affect the signatures of observed discharge for gauging stations across Europe, (b) whether or not the model reproduces these relationships, (c) whether or not the model can reproduce the spatial variability in runoff across Europe, and (d)

which aspects of discharge the model can sufficiently reproduce for catchments of varying scales and characteristics across Europe.

2 Data and methods

The study uses the HYdrological Predictions for the Environment (HYPE) distributed hydrological model (Lindström *et al.* 2010), which was set up with a daily time-step over Europe (i.e. E-HYPE_v2.1) for 35 447 sub-basins with a median size of 214 km^2. The model application is entirely based on open and readily available data sources. The model results were evaluated for the period 1981–2000 in 181 river gauges, which had not been used for model calibration (with the exception of lake rating curves).

2.1 Discharge data and flow signatures

A database of observed discharge data was created for establishing flow signatures, using data downloaded from the Global Runoff Data Centre (GRDC), the European Water Archive (EWA) and the Baltex Hydrological Data Centre (BHDC) databases, initially including 1200 stations (Table 1). Data were then filtered for catchment area greater than 5000 km^2, after accounting for their availability in the period 1981 to 2000, as well as removing stations with large portions of missing data and highly unusual hydrographs (i.e. measurement and reporting errors, or new dams or regulation schemes). The latter check was done manually by visually inspecting each hydrograph. The limit of 5000 km^2 was used because it was shown by Donnelly *et al.* (2013a) that, for catchments smaller than this, errors in catchment delineation are frequent in the open database used (i.e. HydroSheds). Finally, nested gauging stations (in shared catchment areas) were removed, so that each gauging station represents at least 90% unique upstream area. This was done to avoid bias in results to those river catchments where most data are available.

After filtering, a total of 181 river gauges remained with either daily (157 stations) or monthly (24 stations) time series, providing 20 years of observed river discharge across continental Europe (Fig. 2). These sites were well distributed over the model domain with the exception of France, Italy, Eastern Europe north of the Black Sea, and in Turkey, where data were lacking. Of the 181 selected river gauges, 132 were regulated by upstream dams and reservoirs to the extent that this could be identified visually in the hydrographs. The remaining catchments may still have had smaller amounts of regulation but this information was not available for all the gauges.

Six flow signatures were calculated for each gauging station to illustrate different aspects of the streamflow variability and the hydrograph characteristics (Appendix A). The signatures were: the long-term mean river discharge (MQ), the mean specific discharge (MQ/A, normalized by catchment area), the normalized 95th percentile representing high flows (Q_{95}/A), the normalized 30th percentile representing low flows (Q_{30}/A), the coefficient of variation over time representing the dominant flow variability (CV), and the hydrograph flashiness representing the daily flow variability (Flash).

Table 1. Open data used in the database of catchment physiography and set-up of E-HYPEv_2.1. Full references are given in Table S1 in the Supplementary Material (numbers in square brackets).

Variables	Detail/resolution	Data source
Areal extent	8.8 million km^2	Hydrosheds [1] and Hydro 1K (for latitude > 60°) [2]
Median sub-basin resolution	214 km^2	Tailored for E-HYPE2.1
No. sub-basins	35 447	Tailored for E-HYPE2.1
Topography and routing	15 arc-second	Hydrosheds [1] and Hydro 1K (for latitude > 60°) [2]
Temperature and precipitation	0.75 degree	ERA-INTERIM [3] corrected to Global Precipitation Climatology Centre (GPCC) for monthly mean precipitation [4]
Land cover	9 classes defined	CORINE [5] and GLC2000 (for areas not covered by CORINE) [6] Euroland Soil Sealing 2009 [7] for urban area GLWD [8] for lake area and distribution
Lake and reservoir information	2.2% of domain	GLWD (Global Lake and Wetland database) [8] ERMOBST [9] FLAKE-Global [10] International Water Power & Dam Construction yearbook 2011 [11] ILEC World Lake database[12] LEGOS [13] Swedish Water archive (SMHI) [14]
Agriculture		CAPRI [15] MIRCA2000[16], FAO-56 [17]
Irrigation	16% of domain	EIM, EU scale [18] GMIA, global scale [19] Siebert *et al.* (2010), global scale [20] FAO-56, regional scale [17]
Soil types	4 soil classes	For catchment physiography: ESD [21] and DSMW [22] For E-HYPE 2.1: based on land use
River discharge	1200 stations	GRDC [23], EWA[24], BHDC[25]

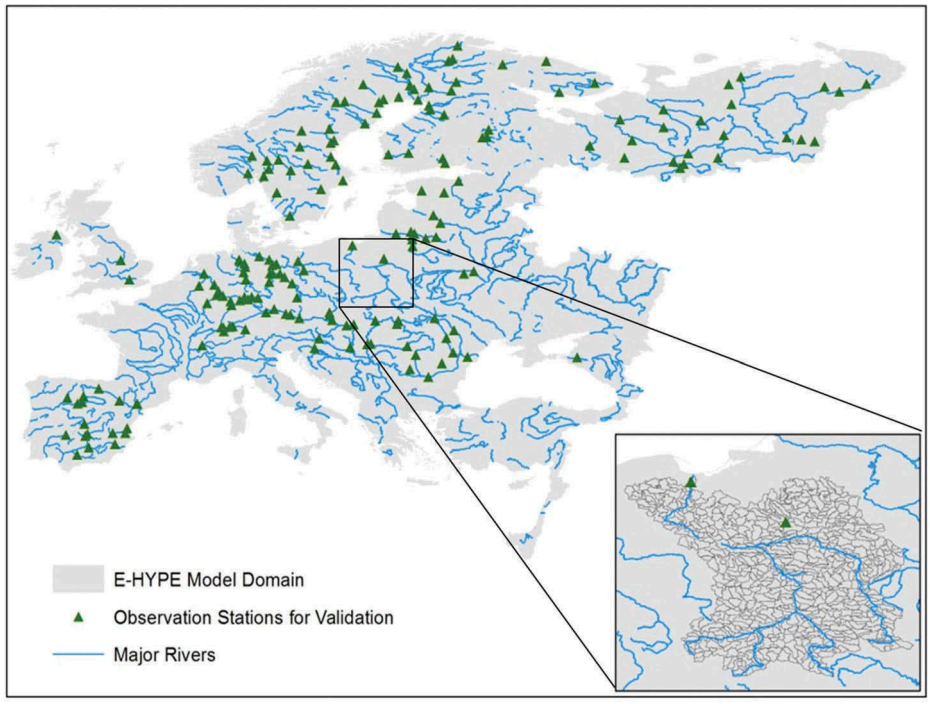


Figure 2. Domain of continental Europe, major rivers and gauges used in the study. The inset illustrates the resolution of sub-basins for the Vistula River (bottom right).

2.2 Catchment characteristics

Physiographical data were compiled at catchment level for continental Europe, using GIS and data from readily available regional or global open datasets (Table 1). The domain was divided into 35 447 linked sub-basins with a median resolution of 214 km^2 (Fig. 2) based on the HydroSheds database (Lehner *et al.* 2008). Reanalysis temperature and precipitation data for sub-basins were taken from the nearest meteorological grid point to each sub-basin centroid. Data for land cover (e.g. vegetation) and soil type were aggregated into fewer classes than in the original databases. Lakes and reservoirs with over 10 km^2 surface area were explicitly included at the outlet of sub-basins (in total 1955 lakes), while smaller lakes were included as internal lakes (in total 20 279 lakes).

The database was used for setting up the E-HYPE_v2.1 model and for identifying catchment characteristics upstream of 181 gauging stations across Europe with the exception of soil types. Soil types upstream of the gauging stations were originally taken from the European Soils Database (Panagos 2006) and the Digital Soil Map of the World (FAO 1995) for the correlation analysis, but the final model version uses four soil classes based on land use. This was because the correlation analysis as well as ground truth in Sweden showed that the European-scale database was unable to represent sufficient variability in European soil types. It could be seen that the level of detail in the national databases provided to the ESDB varied considerably by country.

Physiographic variables chosen for analysis of correlation with flow signatures were: catchment area (A); mean values for annual precipitation (P), annual temperature (T), gauge latitude (Lat), average catchment elevation (Elev); percentage of catchment area classified as organic soils (S_Org), fine soils (S_Fine), coarse soils (S_Coarse), agricultural land (L_Agric), forested land (L_Forest) and lakes (L_Lake).

2.3 The E-HYPE_v2.1 model

The HYPE code is a semi-distributed and process-based model using conceptual algorithms in sub-basins and along flow paths (Lindström *et al.* 2010). When applied across Europe, it is called E-HYPE and the current version is v2.1. The model is forced by daily precipitation and temperature and then calculates flow paths in the soil based on snow melt, evapotranspiration, surface runoff, infiltration, percolation, macropore flow, tile drainage, and lateral outflow to the stream from soil layers with water content above field capacity. Calculations start for the Hydrological Response Units (HRUs) in each sub-basin. The HRUs in HYPE reflect a unique combination of soil type and land cover, which are assumed to govern the flow paths and discharge from soils. A HRU has a set of parameters specific for the soil type/land cover combination,

which is not catchment-specific but distributed all over the model domain. Each sub-basin may consist of any combination of HRU types and each HRU can be divided vertically into three distinct soil layers (normally top-soil, root zone and the remaining soil are distinguished). HRUs are connected directly to the stream and act in parallel. The groundwater level in each HRU is fluctuating, may saturate the soil layers and water may percolate between sub-basins. Groundwater discharge is pressure dependent and, along with the outflow from the soil layers, it is then routed within and between sub-basins using a river routing routine, which simulates attenuation and delay in rivers, and storage in lakes. In contrast to HRUs, the parameters describing delay in lakes and regulation of dams may be catchment specific or general, while parameters describing snow processes, percolation between sub-basins or delay in rivers may be either general or regionalized. For the E-HYPE_v2.1 set-up, these parameters were general, i.e. consistent across the domain, but lakes and reservoirs with downstream gauging stations were calibrated individually.

The influence of explicit lakes at the sub-basin outlet is modelled by a rating curve, which may be generalized for many lakes or specified individually for a given lake. In addition to natural dynamics, which are described by a rating curve, the model can simulate simplified water management schemes. Reservoirs or regulated lakes can have a specified target release which is either constant or seasonally variable, as is often the case with hydropower. This is calibrated based on downstream gauge data where available. The other major anthropogenic intervention that is included is irrigation. Irrigation is simulated based on crop water demands calculated either (i) with the FAO-56 crop coefficient method (Allen *et al.* 1998) or (ii) relative to a reference flooding level for rice. The demands are withdrawn from rivers, lakes, reservoirs, and/or groundwater within and/or external to the sub-basin where the demands originated. The demands are constrained by the water availability at these sources. After subtraction of conveyance losses, the withdrawn water is applied as additional infiltration to the irrigated soils from which the demands originated.

The HYPE model has several parameters, for which the modeller has to set values in the model set-up procedure and normally these are constrained in a stepwise manner using different types of observed data (Arheimer and Lindström 2013). The parameters may be soil type dependent (e.g. field capacity), land cover dependent (e.g. evapotranspiration coefficient) or general across the domain (e.g. river routing parameters). Parameters for a HRU have no connection to locality or catchment, but are specific for a given soil type/land cover combination and are then applied anywhere in the domain where that combination of soil type/land cover exists. Lake rating curves and seasonal regulation may also be calibrated to the nearest downstream gauging station, if available. The HYPE model is continuously developed and available in several versions for free download at http://hype.sourceforge.net/. In this study we have used the HYPE code version 4.1.0. Full details on the process descriptions are available online in the code documentation for this version.

The first E-HYPE model was set up in 2008 (Donnelly *et al.* 2009) using the databases available at that time (Strömqvist *et al.* 2009). Not only the code but also the model set-ups are version managed at SMHI, and in this paper we used the E-HYPE model v2.1 based on the database described above (Table 1). It was run on a daily time-step, using the ERA-INTERIM reanalysis at 0.75 degrees (Dee *et al.* 2011), which was adjusted to match the monthly climatological precipitation means from the GPCC database (Rudolf and Schneider 2005) at 0.5 degrees. We used a stepwise approach for estimating different groups of model parameter values in each step:

(1) For evapotranspiration, a general potential evapotranspiration parameter was manually optimized aiming to minimize volumetric errors relative to observed data from 20 flux towers from European Fluxes Database Cluster (http://www.europe-fluxdata.eu/), Fluxdata.org (http://www.fluxdata.org), and Consortium for Small Scale Modelling (http://www.cosmo-model.org/), all accessed in September 2011.
(2) Initial soil and land cover parameter values were taken from a calibrated parameter set for Sweden (Strömqvist *et al.* 2012) and were then tuned by optimizing parameters to groups of representative gauged basins (RGBs) for each soil type and land cover. In all, 104 gauged catchments larger than 1100 km^2 were used to calibrate the soil and land cover parameters. The Nash-Sutcliffe Efficiency (NSE, Nash and Sutcliffe 1970) was used as an objective function for each group of RGBs. For example, parameters for the land cover forest were optimized simultaneously to 24 small, lake-free gauged basins across the domain for which forest was the dominant land cover (>70%).
(3) General catchment parameters, i.e. parameters that are constant across the model domain, but not linked to soil type or land cover, were calibrated to optimize the model performance across the full group of calibration stations. This included further

tuning of the general potential evapotranspiration parameter to minimize the error in simulated volume.

(4) Individual rating curves for lakes were calibrated for sites where gauges were co-located with lakes (i.e. for the 121 largest lakes). Individual regulation schemes (target releases) were fitted to describe the functioning of 46 major reservoirs across Europe. Other lakes and reservoirs in the model where simulated using a general rating curve.

The agriculture and irrigation datasets listed in Table 1 were used to define irrigated area, crop types, growing seasons, crop coefficients, irrigation methods and efficiencies, and irrigation sources. Regional irrigation transfer networks were estimated based on proximity to relatively large dams within the catchment. The regulation for some of these dams was adjusted to fit the irrigation season. The irrigation parameters regulating water demand and abstraction were manually calibrated using discharge stations in irrigation-dominated areas with the Kling-Gupta Efficiency, KGE (Gupta *et al.* 2009). The KGE was used here because volume errors in irrigated areas were sometimes high, and analysing the KGE components (correlation, mean ratio, variability ratio) enabled reasonable calibration also in these areas.

2.4 E-HYPE_v2.1 model evaluation in blind tests

Continuous time series of water discharge modelled with E-HYPE_v2.1 were evaluated for the 181 selected river gauges, which represent more heterogeneous land cover and soil type conditions than the representative gauged basins used for calibration (step 2 above). The model was thus evaluated for catchments with varying degrees of anthropogenic alteration, as these are conditions that the model needs to replicate. Although there are probably no completely pristine catchments larger than 5000 km^2 in Europe, natural flow dominates the hydrographs of many catchments. Evaluation of the ability of the model to capture spatial and temporal variability in discharge was done by comparing modelled flow signatures, interannual, seasonal and daily simulations with observed data. Several statistical criteria were used in the evaluation (Appendix B): relative volume error (RE), spatial correlation of mean annual discharge and temporal correlation of annual discharge and explained variance of daily discharge using NSE and KGE. In addition, seasonal discharge and relative errors of flow percentiles were plotted.

2.5 Relationship between flow signatures and catchment characteristics

In addition to the traditional model performance metrics (above), catchment characteristics were correlated with flow signatures in observed and modelled time series to determine: (a) which catchment characteristics affect discharge; (b) whether the HYPE model assumptions on catchment similarities (e.g. definition of HRUs) are valid at this scale and for the chosen input data; and (c) whether the E-HYPE_v2.1 model captures the same dynamics as can be observed in field data. For the correlation matrix, we used distance correlation, dCor, to assess the dependence between the variables (Székely *et al.* 2007). Distance correlation was chosen to capture non-monotonic associations and ensure that the correlation is zero if and only if the variables are statistically independent, which is not the case for, e.g. Pearson's product-moment correlation. The range of dCor is 0 to 1; and the closer dCor is to 1, the stronger the dependence between the variables. The statistical significance of the dependence was assessed using the distance covariance test with 1000 replicates at the 1% significance level (Székely *et al.* 2007). Finally, we directly compared the observed and modelled flow signatures in order to evaluate the model's capacity to simulate these in ungauged basins.

3 Results

3.1 Correlations between flow signatures and catchment characteristics

Results from extracting flow signatures and catchment characteristics for the 181 stations show that there is a significant spread among the studied catchments (Table 2). Significant correlations to physiography could be found for each of the six flow signatures studied (Fig. 3, upper right quadrants). For observed flow, however, some of these correlations were relatively weak (Fig. 3(a)), explaining less than 50% of the variation, while stronger correlations were found for the modelled flow (Fig. 3(b)).

As expected, the magnitude of mean river flow (MQ) was mostly related to catchment size (A), but also to all other catchment characteristics studied and among them most strongly to precipitation (P). This was found for observations (Fig. 3(a)) as well as for modelled data (Fig. 3(b)) and is reflected in the structure of all rainfall–runoff models that use the precipitation, P, falling over a catchment, A, as inputs for simulating runoff. When removing the influence of area for discharge and studying high, mean and low flows (e.g. Q_{95}/A, MQ/A and Q_{30}/A), stronger

Table 2. Distribution of flow signatures and catchment characteristics of the 181 river gauges based on 20 years of recorded daily time series and the E-HYPE_v2.1 database. See Appendix A for definitions of the flow signatures.

Variable	Description	MIN	MAX	MEAN
Flow signatures:				
MQ	Long-term mean discharge (m^3/s)	2	6293	453
MQ/*A*	Mean specific discharge (mm/year)	3	1201	354
Q_{95}/A	Area-weighted high flow (95th percentile) (mm/year)	6	3872	924
Q_{30}/A	Area-weighted low flow (30th percentile) (mm/year)	0.4	801	192
CV	Coefficient of variation of the daily discharge time series (–)	0.33	4.1	0.88
Flash	Daily hydrograph flashiness (–)	1.6×10^{-3}	0.25	4.2×10^{-2}
Physiography: *A*	Catchment area (km^2)	3988	807 000	42 880
Elev	Mean elevation of catchment (m a.s.l.)	61	1745	460
Lat	Latitude of sub-basin outlet	37.4	70.1	54.8
L_Lake	Percentage lake in catchment	0%	20%	3%
L_Forest	Percentage forested area in catchment	2%	98%	49%
L_Agric	Percentage agricultural area in catchment	0%	85%	24%
S_Coarse	Percentage coarse soils in catchment	0%	92%	23%
S_Fine	Percentage fine soils in catchment	0%	66%	6%
S_Org	Percentage organic soils in catchment	0%	59%	4%
T	Mean temperature of catchment (degrees C)	−7.4	16.0	5.5
P	Mean precipitation over catchment (mm/year)	434	1895	751

correlations to other catchment characteristics were found. Both *P* and elevation (Elev) were more strongly correlated to high annual flow when area weighted and highly correlated to Q_{30}/A and Q_{90}/A. The results for high flows (Q_{95}/A) were similar to the results for the mean flows, but with higher correlation to temperature. This is probably due to both the pronounced snow-melt peak in cold regions and more flashy floods in warm/arid regions. Looking at the scatter plots (Fig. 3, lower left quadrants) higher peaks are found in catchments with both lower and higher mean temperatures. The model reproduces a temperature relation weaker than that observed, but is more strongly linked to precipitation. The temperature (*T*) was negatively correlated, probably as loss through evapotranspiration increases with higher temperature.

There was also an indication that land cover and soil type affected flow signatures. Agricultural land (L_Agric) was strongly negatively correlated to Q_{95}/A, Q_{30}/A and MQ/*A*, but also positively to Lat and CV and negatively to *T*. The correlation of agriculture with low base flow may be due to the use of clay soil for agriculture for which base flows are often low (Arheimer and Lindström 2013), but may also reflect the removal of water for crop requirements. However, there is a north/south trend in agricultural land in Europe with more agriculture in warmer, drier, more southerly catchments, so it cannot be ruled out that the negative correlation of agriculture with flow simply reflects the climate of agricultural regions in Europe. The flow signatures describing the temporal variability of the flow (i.e. Flash and CV) also show some significant but weak correlations to catchment characteristics. Lakes have a rather strong damping effect on the observed hydrographs (reducing the CV), yet not as strong as in the modelled data.

Low flows (Q_{30}/A) were positively correlated to lakes, coarse soils, elevation and precipitation, indicating that these factors result in higher base flow. Lakes have a damping effect on water variability and dry periods. Moreover, the results indicate the ability of wetter catchments to maintain soil moisture, store precipitation and thus increase the base flow. Nevertheless, the relationships between flow signatures and catchment characteristics were very similar when using observed and modelled time series, respectively (e.g. Figure 3(a) and (b)). Similar correlations were found for lakes, forest, precipitation and temperature. Both observed and simulated flow signatures were significantly but weakly correlated to the flow signatures for fine soils with stronger relationships in the observed data for Q_{95}/A, Q_{30}/A and MQ/*A* but weaker for CV and Flash. Latitude and temperature were more strongly correlated to observed high flows than to modelled high flows, and may be explained by the insufficient resolution of the precipitation data in the modelling. In contrast, the model showed stronger correlation between catchment characteristics and flashiness of flow than was found in observed data. This suggests factors other than physiography affect the daily variability in river flow (e.g. human alterations) and that these are not yet adequately represented in the model. The discrepancies found will provide direction for future refinements of the model.

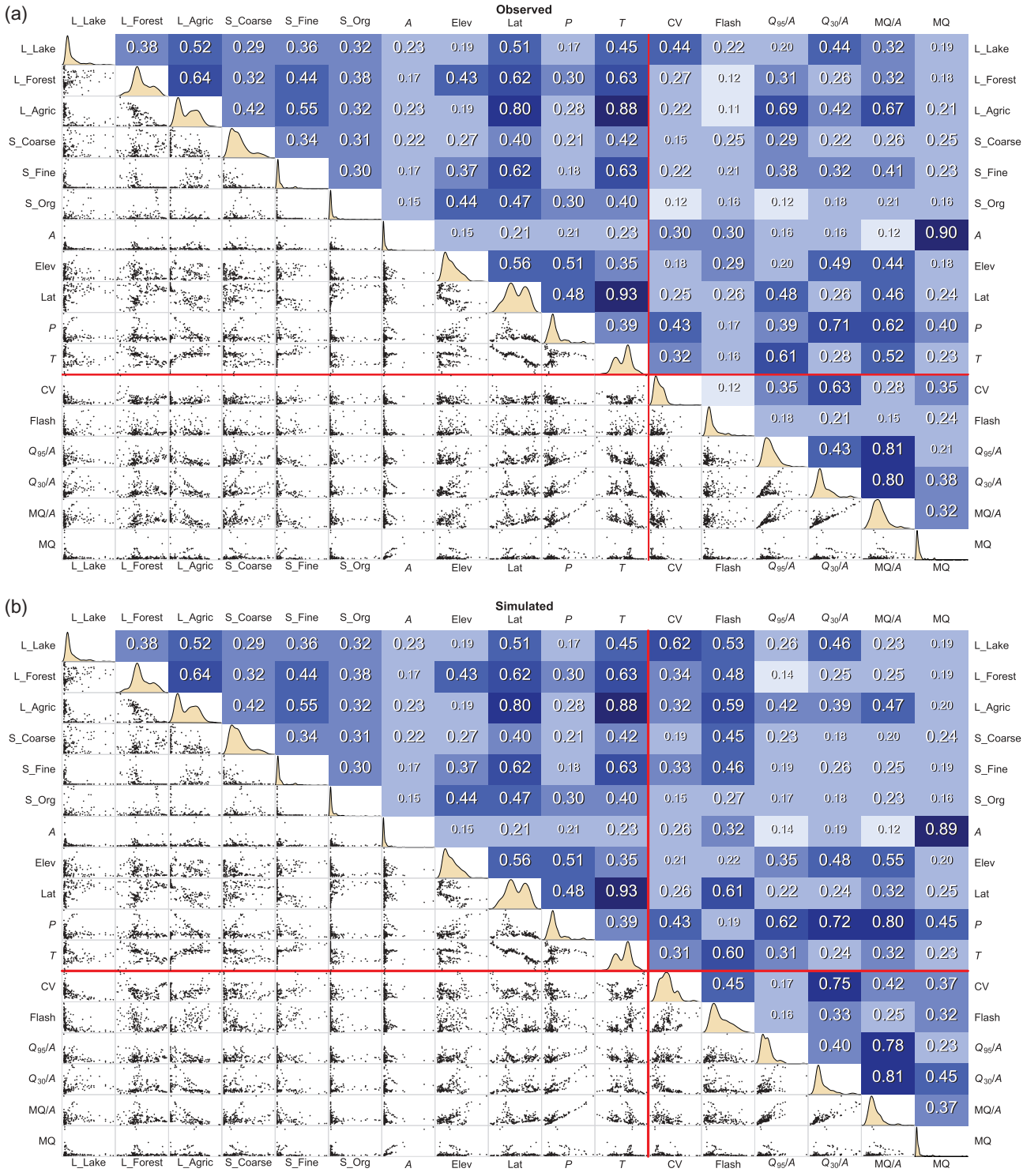


Figure 3. Correlation (dCor) matrix of catchment characteristics and (a) observed flow signatures, (b) simulated flow signatures, using E-HYPE_v2.1. The dCor values in larger print are statistically significant at the 1% level. Darker blue/shading highlights stronger correlation (dCor closer to 1, the maximum), and lighter blue indicates weaker correlation (dCor closer to 0, the minimum). See Table 2 and Appendix A for definition of the terms. The red/thick lines define four quadrants for ease of interpretation. Note that the upper-left quadrant is the same in (a) and (b) (correlating catchment characteristics on their own), while the other quadrants differ (correlating flow signatures on their own, and flow signatures with catchment characteristics, respectively).

3.2 Multi-criteria evaluation of E-HYPE_v2.1 in validation sites

3.2.1 Model predictions of flow signatures

The similarity of the observed and simulated flow signatures shows that the E-HYPE_v2.1 model captures rather well the spatial variation across Europe in long-term averages of discharge (MQ, MQ/A), low flow (Q_{30}/A) and high flow (Q_{95}/A) (Fig. 4). Given the high correlation of these signatures with precipitation, this reflects the response to spatial variation in the input precipitation. Considering all flow signatures simultaneously in a multi-dimensional distance correlation, this analysis suggests that the model is very similar to the observations (dCor = 0.98), even if one removes the effect of MQ (dCor = 0.76). The dominant temporal variability of the flows is also captured well (CV; r = 0.75), suggesting that the model does capture monthly and inter-annual variability at most locations. The daily variability of the hydrographs (Flash) was more difficult to capture, especially for catchments with large daily variability (i.e. high observed values of Flash), where the model results were poor. These catchments are normally found in mountains and in the Mediterranean region, which can have very local rain events resulting in flash floods. This process is not sufficiently represented in the model as the precipitation is estimated from a 0.75 degree grid.

Figure 5(a) shows that the average river flow for about 32% of the validation catchments could be estimated with less than 10% relative error (RE) and 75% of catchments with less than 25% RE. The remaining 25% of stations had REs exceeding 25%. There was a negative bias in discharge for northern Europe, which is probably caused by underestimated precipitation in the forcing dataset. When comparing the forcing data for Swedish catchments with the Swedish national dataset (PTHBV; Johansson 2002), there was an average 10% underestimation of precipitation in Sweden. This is because the GPCC dataset includes insufficient stations in Sweden (and other northern countries), particularly at higher elevations and does not include any corrections for undercatch, which is significant in the snowier northern regions. The stations showing overestimation in flow are mainly located in southern Europe. In some cases we know this to be caused by underestimated recharge to depleted aquifers which is not simulated by the model. Insufficient representation of irrigation, domestic use and industrial extractions in the model may also contribute to the results for this part of Europe, which is highly populated and its hydrology is strongly affected by human activities.

3.2.2 Model predictions of temporal variations

Although the model output is daily, the model's ability to simulate different temporal scales of discharge is evaluated because multi-basin models are often used to deliver simulated data for many different purposes. The ability of the model to reproduce the inter-annual variability of discharge was assessed by calculating Pearson's correlation coefficient (r) between simulated and observed mean flow for each year at each gauging station. Inter-annual variability indicates the model's

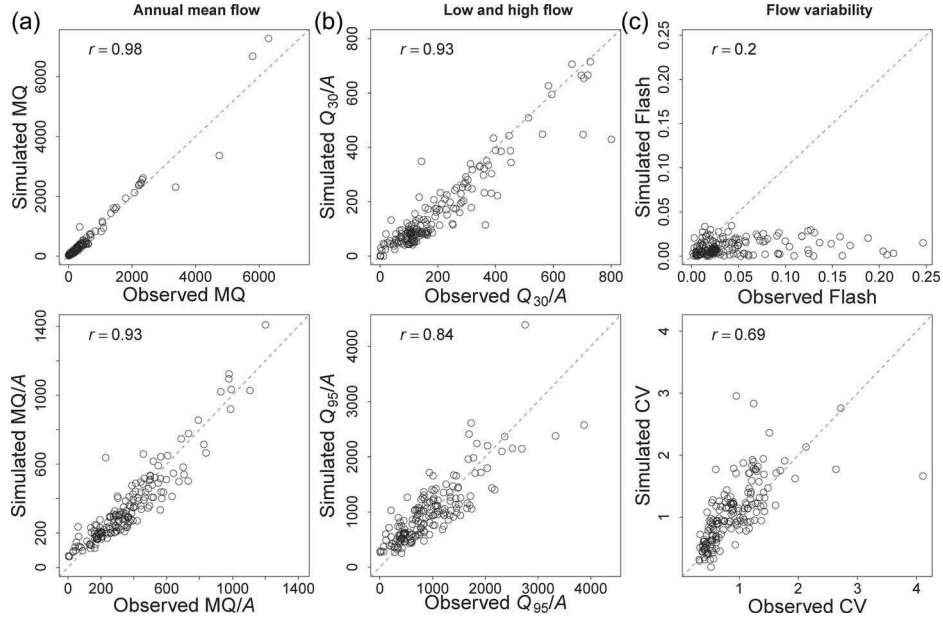


Figure 4. Scatterplots of simulated *vs* observed flow signatures, describing: (a) annual mean flow, (b) low and high flow, and (c) temporal flow variability at each gauging site, where r is Pearson's correlation coefficient.

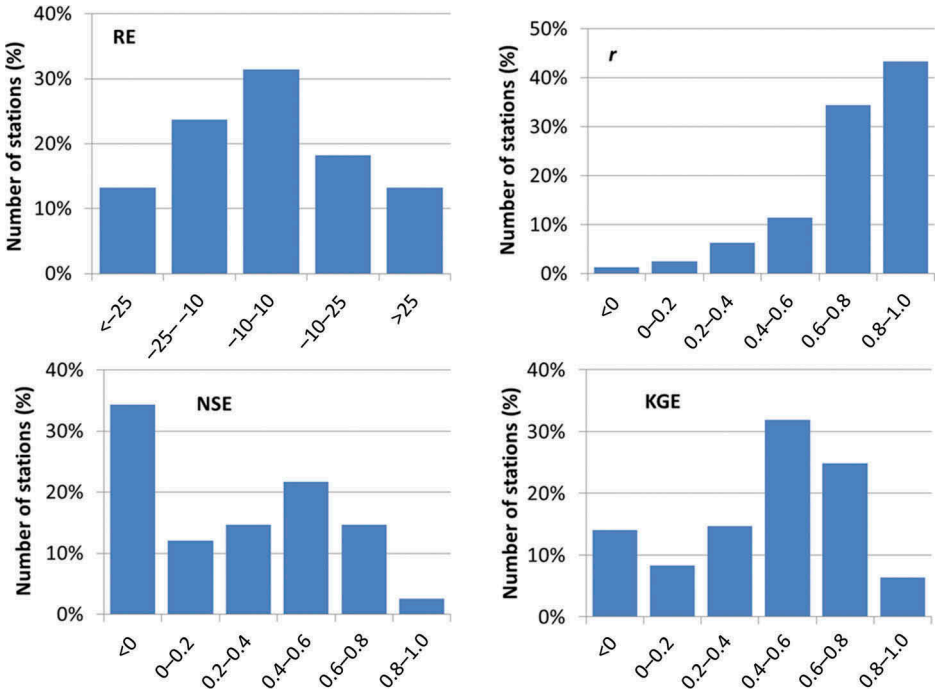


Figure 5. Histograms of E-HYPE_v2.1 model performance for different measures at 181 validation stations (157 for daily data) across Europe (for locations, see Fig. 2).

ability to respond differently to wet and dry years, for example. More than half of the stations had correlation coefficients exceeding 0.8 and nearly all exceeding 0.4 (Fig. 5(b)), indicating that the model responds well to inter-annual variation in precipitation and temperature. There were some regional patterns in the correlation coefficients, with poorer correlation results on the Iberian Peninsula, Northern Sweden and Finland, and in Eastern Europe. This could possibly be due to the model's inability to capture regulation in very large reservoirs which redistribute discharge across years for hydropower or irrigation. For the Iberian Peninsula, the problem is again believed to be caused by underestimated recharge.

The analysis of NSE and KGE across all stations indicates the model's ability to capture daily variations in discharge. For all validation stations across Europe, two-thirds had NSE greater than 0, which means that the model added value compared to using a long-term mean value. More than one-third of the stations had NSE greater than 0.4. KGE is a more useful tool than NSE for comparing performance across catchments of varying seasonality and magnitude. For all validation stations across Europe, 85% had KGE greater than zero and 52% greater than 0.5.

The ability of the model to reproduce the daily variability of discharge varies significantly regionally, as shown by the NSE and KGE (Fig. 5(c) and (d), Fig. 6). The NSE results were better in unregulated rivers than in regulated ones, and were improved with increasing catchment size. This was expected as regulation strategies are non-systematic and difficult to reproduce, and large rivers are less peaky, with compensating upstream model errors giving higher model performance at the larger scale. Analysing the components of KGE, correlation (r), variability ratio (α) and mean ratio (β) reveals that the model errors have different causes. For example 73% of the stations had $0.5 < \alpha < 1.5$, while 92% had $0.5 < \beta < 1.5$, thus indicating that problems with simulating variability contribute more to low KGE than problems simulating volume. Regionally, the KGE components discern different issues at different locations (Fig. 6). In southern Germany and Austria, KGE is low primarily due to α, while β and r are quite good. Hence, the model overestimates the magnitude of the variability, while it satisfactorily simulates the average flow and the relative magnitude on different days. For example, for the Linz station on the Danube River, KGE = 0.16, with $r = 0.76$, $\beta = 1.19$ and $\alpha = 1.78$. Problems simulating variability here may be related to insufficient simulation of flow regulation for navigation purposes. Northern Russian stations also have good correlations, and errors are either caused by underestimation of the means or the variability, or overestimation of the variability, for example at the Ust-Pinega station on the Severnaya Dvina River, KGE = 0.65, with $r = 0.94$,

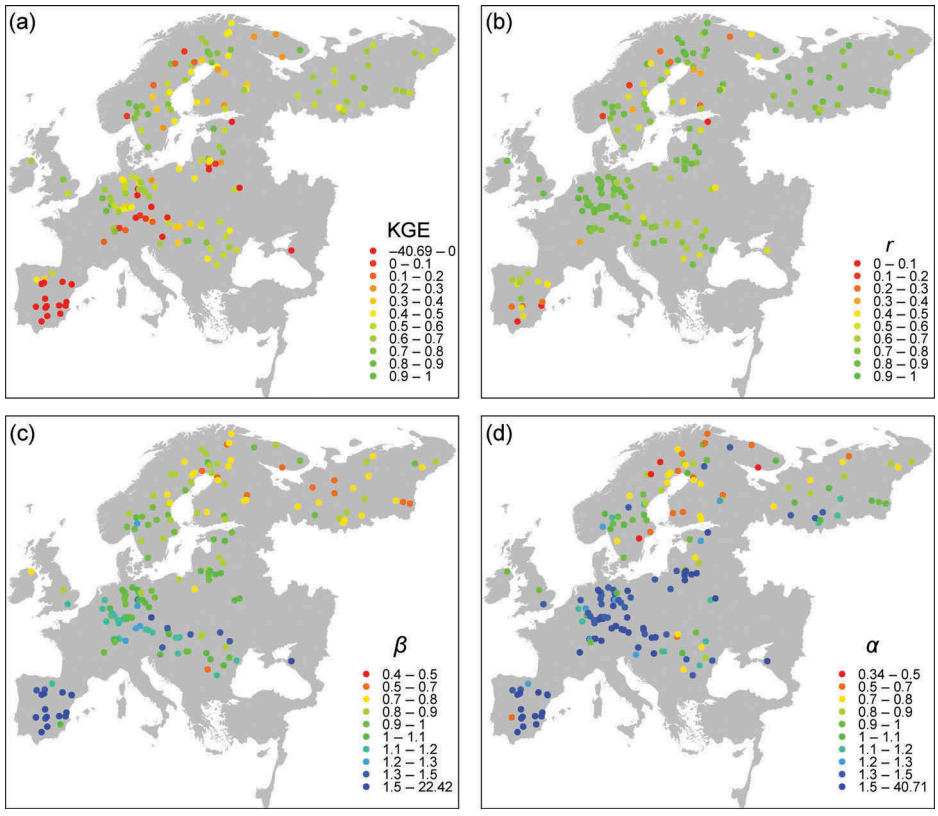


Figure 6. Spatial distribution of the Kling-Gupta efficiency and its components: (a) KGE, (b) correlation (r), (c) mean ratio (β), and (d) variability ratio (α).

β = 0.69 and α = 0.84. Hence, the error in the mean is the major problem at Ust-Pinega. Worse performance (KGE = − 1.63) was obtained for the Tikhovsky station on the Kuban River (Russia, near the Black Sea). Here the model fails because of errors in both mean and variability (r = 0.59, β = 2.75, α = 2.91). Several reasons may cause this failure, e.g. inaccurate representation of irrigation and other abstractions, incorrect catchment delineation or precipitation errors. At the Chancy station on the Rhone River, the mean is reproduced but the variability and correlation are problematic (KGE = 0.15, with r = 0.32, β = 0.97, α = 1.51). Errors in reproducing flow in the Scandinavian mountains are of various origins. At the Storsjön station on the Indalsälven River, the mean flow (0.86) and the magnitude of the variability (0.74) are somewhat underestimated but still acceptable. Instead, it is the inability of the model to simulate timing of flow increases and decreases correctly (r = 0.19) that particularly reduces performance (KGE = 0.13). This is mainly because the model can only reproduce the seasonal cycle of regulation and not the daily regulation. In contrast, the model performance at the Storavan station on the Skellefteån River (KGE = 0.11) is mainly affected by substantial underestimation of the variability (α = 0.34, β = 0.89, r = 0.41), possibly due to an overestimation of the damping effects of lakes, as seen in the correlation analysis.

For further evaluation, the E-HYPE_v2.1 model skill in estimating mean monthly discharge cycles was evaluated for river mouths with direct discharge to the major seas around Europe, as this information is, for example, of interest when defining inflows to seas. Seasonality is generally well reproduced by E-HYPE_v2.1, except for the Mediterranean region where the model exaggerates the winter flow (Fig. 7). For the rest of the continent, the differences in seasonal flow patterns between Northern and Southern Europe are well reproduced. However, again it can be noted that the model underestimates the river flows from the north and overestimates the river flows from the south. The best model results are achieved for rivers with discharge to the European Atlantic coast.

A further evaluation was made to assess the model's ability to simulate different flow percentiles. Figure 8 shows that the spread in model error increases significantly for flows below the 30th and above the 80th percentile. There is also a trend in the model's ability to simulate different flow percentiles, with the lowest flows at each station more often underestimated, while the highest flows are more often overestimated. There are many factors that could cause this. Overestimation

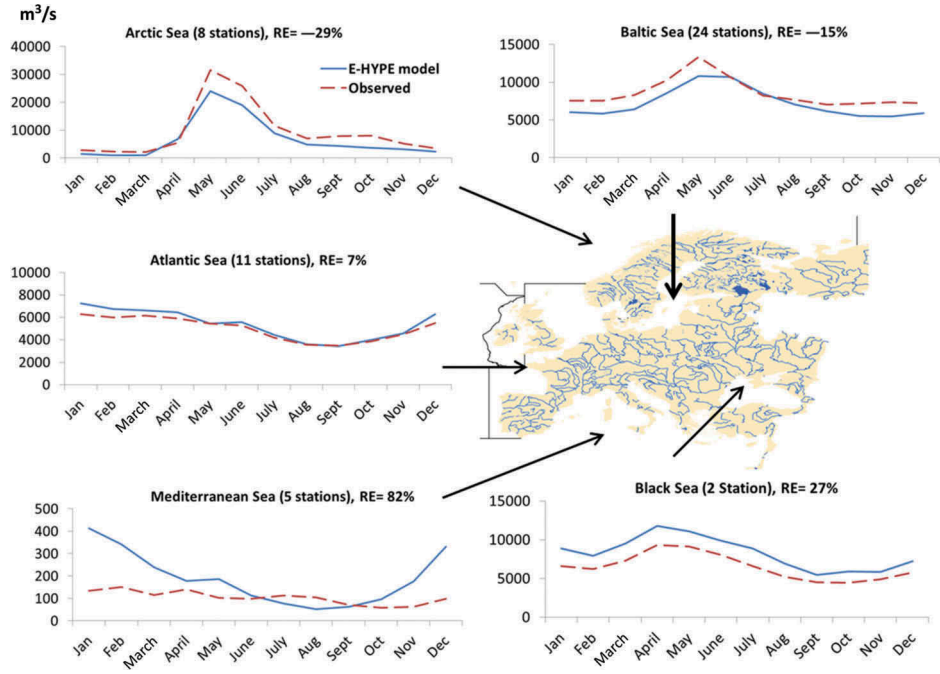


Figure 7. Observed and modelled average monthly inflow cycles (1981–2000) to the different seas surrounding Europe at gauging stations in some major river mouths.

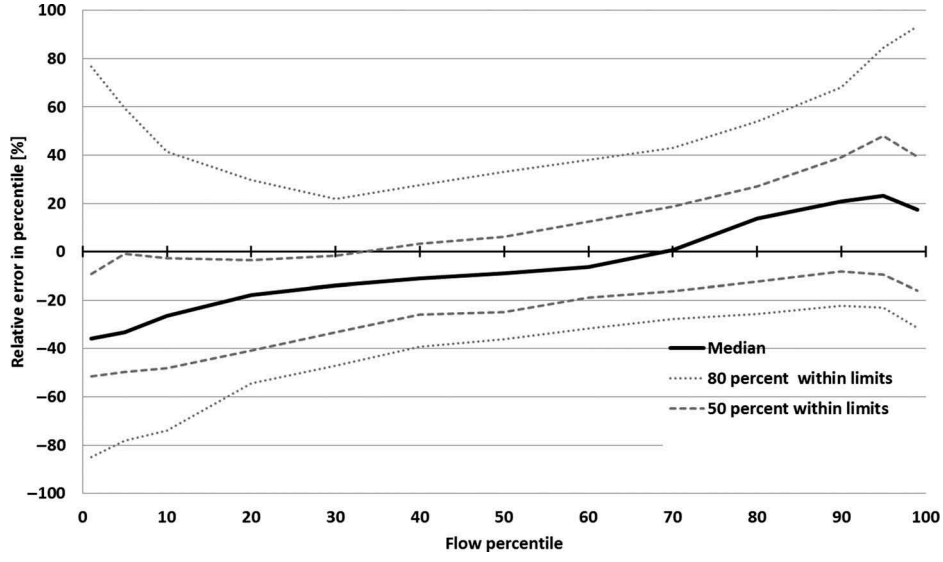


Figure 8. Relative volume error (RE) for flow percentiles in E-HYPE_v2.1.

of high flows may be due to a model parameter that underestimates storage of soil moisture; also, in regulated rivers and rivers with natural lakes, retention may be underestimated. Low flows are difficult to simulate in Europe due to recharge to deep groundwater and the level of anthropogenic influence, including extractions for irrigation, domestic and industrial uses and power station cooling, regulation of rivers for irrigation, flood control and hydropower and diversion via canals. Although the model has simple routines for simulation of irrigation regulation and extractions and hydropower regulation, these routines can only simulate an assumed pattern of human behaviour. There is significant scope for improving the model's ability to simulate low flows.

4 Discussion

Through the above analyses we demonstrated that multi-basin models are useful tools in comparative

hydrology, both for exploring basic assumptions and present knowledge of flow variability in many basins of varying character and size across Europe, and for providing useful hydrological information to a variety of users. These concepts are discussed hereafter.

4.1 Evaluation of model structure using comparative hydrology

Relationships between flow signatures and catchment characteristics were compared for both observed and simulated flow to determine whether the model structure includes those processes important for reproducing the wide variability in hydrology across a continent. Previous studies also show some relationships between flow signatures and catchment characteristics in small, pristine basins, based on large samples of basins (e.g. in the context of MOPEX data; Andréassian *et al.* 2006). For instance, Sawicz *et al.* (2011) searched for relationships between flow signatures and catchment characteristics for 280 basins less than 10 000 km^2 across the eastern United States. They found that spatial proximity caused catchment response similarity and that permeable soils buffered catchment response to climate. With E-HYPE_v2.1 we test the concept using readily available continental-scale data for a large range of catchment scales, and extend the concept to simulated flow signatures, similar to the observed ones.

The E-HYPE_v2.1 model uses the concept of catchment similarity for defining HRUs, so that soil type and land cover are assumed to govern subsurface flow paths and process rates in the model. Once defined, a HRU is simulated anywhere in the model domain using the calibrated parameters. Significant correlations between the observed flow signatures and the catchment characteristics used in the HYPE concept confirm that this is an appropriate assumption. The relationships between simulated flow signatures and catchment characteristics were then used to evaluate whether a single model structure can simulate the spatial variability of discharge across a large heterogeneous region. The results show that the E-HYPE_v2.1 model was able to reproduce five out of six observed flow signatures, and about the same correlation patterns to catchment characteristics as can be found by using observed data. The basic assumption on catchment similarity, which was used in the model concept and for estimating parameter values, is thus quite robust. However, the characteristics for which observed and simulated correlations disagree, indicate potential for model structure improvements (e.g. the damping impact of lakes) or sub-basin specific parameterization transferred using functions of catchment characteristics.

The results indicate that the E-HYPE_v2.1 model includes the major factors contributing to discharge variability, but sometimes exaggerates the linkage to physiography, e.g. for short-term fluctuations of the hydrograph and the damping impact of lakes. In many model concepts, the flow damping effects of lakes is not always explicitly simulated; compare for instance the VIC model (Liang *et al.* 1994). The present E-HYPE_v2.1 model is fairly successful at representing explicit lake routing, but emphasizes this effect a bit too much when extrapolated to overall European conditions, which should be revised in the next model versions.

Correlations of catchment characteristics to flow signatures reflecting short-term fluctuations of the hydrograph were statistically significant but low. It is thus essential to include human impacts and water uses among the catchment characteristics in order to find explaining variables for such patterns, and methods for this are beginning to emerge (Nalbantis *et al.* 2011, Arheimer and Lindström 2014). Precipitation data of higher spatial and temporal resolution would also probably contribute to better explain short-term fluctuations in river flow.

The lack of correlations to some of the soil types, e.g. organic soils, may indicate excess complexity in the HRU descriptions; however, tests with improved input data, particularly regarding soil type, are necessary before abandoning this assumption. Especially, it should be noted that the E-HYPE_v2.1 model is set up to also simulate water quality across Europe (Donnelly *et al.* 2013b) and the HYPE model structure necessarily includes the flow pathways along which nutrients are also transported. Experience with the calibration of water quality parameters in the S-HYPE model (Strömqvist *et al.* 2012) showed that description of correct flow paths for the description of water quality may help constrain parameters describing the quantity part of the model. Hence, the added model complexity used in this study may contribute to model robustness, as calibration improves.

The study demonstrates to what extent the relationships between flow signatures and catchment characteristics can be exploited using readily available data for multi-basin modelling. Several of the assumptions in E-HYPE_v2.1 are used in many other multi-basin hydrological model set-ups using readily available regional and global databases (e.g. WaterGap, Döll *et al.* 2003). Model parameters are assumed constant for a given land cover or soil type, and then optimized simultaneously to many gauged basins and finally

applied via the HRU concept anywhere in the model domain where the soil type/land cover combination of the HRU exists. While there is a compromise in performance at a single gauged basin, average performance across a group of basins is optimized. This approach to modelling could be thought of as the inverse of other regionalization studies, which used flow signatures and catchment characteristics to form multiple model structures and/or parameter sets (e.g. Pechlivanidis *et al.* 2010, Carrillo *et al.* 2011). In these studies, the optimal parameter sets for individual basins were first identified and then the flow signatures were related to the catchment characteristics indirectly (i.e. via the model structure/parameters) allowing for the transfer of parameter sets to other ungauged basins. These approaches allow for the exploration of more precise relationships between flow signatures and catchment characteristics and to produce models with high performance at individually gauged sites; however, to reproduce the same relationship in an ungauged catchment requires the calibration of a relationship to a sufficiently similar set of gauged catchments. In contrast, when using the multi-basin HRU approach, the simulation of HRUs with specific parameter values for each soil type and land cover class means that any combination of HRUs can be simulated anywhere in the model domain. The parameters are determined using regional calibration of many catchments, which has been shown to be the best regionalization approach for ungauged basins based on cross-validation tests (Parajka *et al.* 2007). A further advantage of this distributed approach is that the response of a catchment to land cover changes can be simulated by changing the input HRUs to new land cover and then using the already defined HRU specific parameter values. This makes the distributed approach and process-based model concept useful when exploring interactions between changes in hydrology and society, which is the topic of the IAHS decade *Panta Rhei* (Montanari *et al.* 2013).

4.2 Model performance, usefulness and improvements

The model evaluation considered a number of metrics to judge if the model captured spatial and temporal variability across the large model domain. This approach to model evaluation is intended to assess the usefulness of the model for a variety of purposes. The spatial variability of flow signatures representing long-term mean discharge, low flow and high flow was well reproduced by the E-HYPE_v2.1 model. This makes the model useful for plotting maps of long-term mean conditions at high resolution across the continent and for comparing the relative effects of climate change on such statistics. Temporal analysis shows that the model captures effects of inter-annual flow variations in most stations around the domain. This is a further indicator that the model is suitable for simulating the impacts of climate change on hydrological variables. This should be a standard check for all hydrological models that are used for climate change impact studies at the large scale. Its ability to reproduce the inter-annual variation of discharge also makes the model useful as input to oceanographic models (e.g. Meier *et al.* 2014). There remain, however, modelling issues to improve in the simulation of seasonality in discharge for some regions, particularly rivers draining to the Mediterranean and the Black Sea. The evaluation of daily discharge showed that the majority of catchments could be simulated with satisfactory NSE and KGE, but there remain a large number of catchments for which daily performance is unsatisfactory and only a few sites for which performance could be deemed excellent. Although E-HYPE_v2.1 can represent the spatial variability in high and low flows reasonably, the error analysis of flow percentiles shows that for individual sites the model tends to overestimate floods and underestimate low flows.

These results are comparable to other multi-basin model attempts, where calibration at individual sites is impractical. For example, E-HYPE_v2.1 experiences similar problems in capturing extremes as have been reported for high flows in LISFLOOD (Thielen *et al.* 2009) and for low flows across Europe (Gudmundsson *et al.* 2012). Limitations in performance of multi-basin models have been linked to input data limitations in other studies too (Döll *et al.* 2003, Widén-Nilsson *et al.* 2007). The need for evaluation of model simulations against observed data, particularly regarding discharge from land surface schemes and other multi-basin hydrological models, is well outlined by Stahl *et al.* (2011). Nevertheless, there is a need for a framework for evaluating multiple aspects of large-scale model performance, according to their use. Multi-basin models are often evaluated only for their ability to reproduce a single or a few aspects of temporal performance, e.g. means (Döll *et al.* 2003), means and floods (Hanasaki *et al.* 2008a), non-parametric anomalies (Stahl *et al.* 2011) or single aspects of spatial performance, e.g. only evaluating discharge at outlets of large river basins (e.g. Döll *et al.* 2003, Hanasaki *et al.* 2008b, Hagemann *et al.* 2009), or in very small river basins, below the model resolution (Stahl *et al.* 2011). Given that multi-basin models are often

used to show maps of hydrological variables representing various temporal aspects of flow across the domain, i.e. for basins ranging from the model resolution to the largest of rivers, there is a need to evaluate a model for its ability to represent small, medium and large catchments simultaneously.

The thorough evaluation of model performance presented in this paper also reveals a number of areas for potential improvements in model structure, parameterization and input data. The approach of linking model errors with potentials for improvements allows for gradual accumulation of knowledge as the model structure is updated and improved over time in new versions. For instance, the regional analysis shows catchments in Spain where discharge is systematically overestimated, regions of northern Europe where discharge is systematically underestimated and regions where regulation for hydropower is extensive, e.g. along the upper Rhine and in northern Sweden.

The overestimation of water volume in southern Europe can be linked to the underestimation of irrigation extractions and groundwater recharge. In a previous study for Spain, overestimation in volume was found when not including recharge to deeper aquifers in the region (Lopez-Gunn *et al.* 2011). Moreover, there is significant uncertainty in the irrigation simulations, relying not only on irrigated areas (of acceptable accuracy at European scale, see Siebert *et al.* 2005), but also on uncertain assumptions about irrigation scheduling, efficiencies, conveyance networks, and sources. Often, more withdrawals are made in reality than are reported or than are optimal from a crop sustenance perspective, which would lead to an underestimation of withdrawals in the model and overestimation of discharge (as obtained here). Regarding the underestimation of water volume in northern Europe, this corresponds with regions where snow makes up a significant proportion of precipitation. Even though the ERA-INTERIM forcing data are corrected to GPCC, GPCC does not include corrections for undercatch, which is a significant problem where precipitation falls as snow, nor the underrepresentation of precipitation observations at high altitudes where precipitation is often higher. In the future model versions more attention will be paid to improving precipitation and temperature grids before using them in the model.

The model performance is also generally low in highly regulated rivers. The regulation of reservoirs used for irrigation supply was simulated according to crop water demands. For hydropower, the model code includes the ability to reproduce some aspects of regulation (Arheimer and Lindström 2014) using a target release, but in E-HYPE_v2.1 this is currently only simulated in rivers where a downstream gauge shows the amplitude, phase and shape (sinusoidal or square wave) of the regulation scheme (assuming that this is seasonal). Given that the validation stations were not used for tuning, simulation of regulation is missed in some of these basins, and would likely be missed also in ungauged basins. Furthermore, the simulated regulation scheme only aims to reproduce the seasonality of regulation and when a reservoir spills above its regulation capacity. Higher frequency (e.g. daily or weekly) variation of regulated discharge is not captured by the model, hence the difficulties in achieving high NSE or KGE based on daily values. It is unlikely that daily regulation can ever be reproduced in hydrological models, with gate operation being a complex function of the dam function, current storage levels, energy prices and demand, irrigation need and demand, flood storage, drinking water requirements and more. However, for a large domain, a model that accounts for general reservoir operations may significantly improve the simulations (Hanasaki *et al.* 2006, Arheimer and Lindström 2014).

4.3 Future improvements

The study provides a framework for how catchment similarity can be used in a simple way to evaluate and drive the development of multi-basin models for both operational and research purposes. Although this study indicates how far we can go using existing readily available open data, there are perspectives for future improvements. Large-scale datasets compromise local detail for large-scale coverage. In the input datasets used, there are known problems with precipitation, water divides (Donnelly *et al.* 2013a), soil types in specific regions, etc. Moreover, improvement of process representations could be obtained by extending the number of catchment characteristics and use their combined principal components for finding parameter values (Hundecha *et al.* 2008). Other factors not considered here, but shown to affect flow signatures in other studies (e.g. Merz and Blöschl 2009), include slope, geology and existence of deep groundwater aquifers. In the present model version, slope and soil depth were not included as inputs due to limitations in the regional databases for soils and in defining slope for widely varying catchment scales, but these will be considered for future model versions. At present, there is also an on-going debate on whether information on soil (Samaniego *et al.* 2010) or topography (Savenije 2010, Gao *et al.* 2014) is most efficient for parameter regionalization (e.g. the definition of HRUs). This could be further examined by using the multi-basin

model approach and similarity concept, to compare results based on these various hypotheses.

It should be recognized that large-scale models are dependent on open data for the whole domain and regional estimation of parameter values, which will never be as good as explicit calibration in specific catchments. Nevertheless, the multi-basin approach allows for predictions in ungauged basins and comparative hydrology, which tests hypotheses in a wide range of basins simultaneously. Interesting findings with implications for future model development and model usefulness could be retrieved from the study. Based on our results, we suggest these methods for testing basic model assumptions when applying multi-basin models, and the use of several flow signatures and multiple criteria as standard in the evaluation procedure.

The next E-HYPE model version will be more regionally calibrated using the similarity concept for regionalization of parameter values. Recently, a number of research efforts have been made to develop strategies to relate model parameters to measurable catchment properties using different approaches (Blöschl *et al.* 2013, Hrachowitz *et al.* 2013). Each of these methods has its own advantages and shortcomings, which would be worthwhile to test in a multi-basin model framework.

5 Conclusions

The spatial variability in flow signatures representing long-term mean discharge, low flow, high flow, and the dominant temporal variability could be significantly and strongly correlated to catchment physiography across Europe. Strongest (>50%) relationships were found for upstream area, proportion of upstream agricultural land, elevation, mean precipitation, and mean temperature. The observed flow signatures representing short-term variability of the hydrograph were more weakly correlated, which implies that either shorter term precipitation is of importance or that other factors are then more important, such as human interventions to water resources.

The E-HYPE_v2.1 multi-basin model is useful for predictions in ungauged basins. The model is able to reproduce the spatial variability in runoff generating processes, both for high, mean and low flows across Europe for a wide variety of catchment scales, climates, anthropogenic and physiographic conditions. However, the model has less skill in the Mediterranean and mountainous regions, where the representation of daily hydrographs and extremes is still poor. This implies that the model can be used for providing long-term flow signatures at the annual and seasonal scales, but caution should be taken when using the daily time series or extremes at single sites. We identified knowledge gaps in evapotranspiration, aquifers, resolution of precipitation patterns and water extraction by humans. This will guide the improvements for the next model version.

The study provides a framework for how the similarity concept can be used to drive the development of multi-basin models for operational and research purposes. We suggest using independent gauges and multiple criteria to learn how to improve the model concept and its application. We also suggest evaluating multi-basin models by combining similarity tests, reproduction of flow signatures, seasonality, inter-annual variability, spatial variability, error components, extreme flows, and daily dynamics. Additional evaluation methods could be employed depending on the specific model purpose. Accordingly, the outcome of the model evaluation indicates the usefulness of various model results.

Acknowledgements

We would especially like to acknowledge contributions from Kristina Isberg, Joel Dahné, Johanna Nilsson, Jörgen Rosberg and Lena Strömqvist. In addition, we would like to acknowledge the IHP Friend Network, BHDC and GRDC for providing observed data from river gauges. We thank Maria Rizzo and Gabor Szekely for helpful analytical advice. We thank the two anonymous reviewers and the editors for their valuable comments on the manuscript. The HYPE model code is open source and can be retrieved with manuals at http://hype.sourceforge.net/. Time series and maps from the E-HYPE_v2.1 model are available for inspection and free download at http://hypeweb.smhi.se.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

We are very grateful for funding from the EU FP7 project SWITCH-ON (Grant Agreement no. 603587). The study was performed at the SMHI Hydrological Research unit, where much work is done in common, taking advantage of previous work and several projects running in parallel in the group.

ORCID

Chantal Donnelly http://orcid.org/0000-0002-0086-4453

References

Alcamo, J., *et al.*, 2003. Development and testing of the WaterGAP 2 global model of water use and availability. *Hydrological Sciences Journal*, 48, 317–337. doi:10.1623/hysj.48.3.317.45290

Allen, R.G., *et al.*, 1998. Crop evapotranspiration (guidelines for computing crop water requirements). *FAO Irrigation and Drainage Paper 56*. Rome, Italy: FAO.

Andersson, J.C.M., *et al.*, 2012. Improved SWAT model performance with time-dynamic Voronoi tessellation of climatic input data in Southern Africa. *JAWRA Journal of the American Water Resources Association*, 48, 480–493. doi:10.1111/j.1752-1688.2011.00627.x

Andréassian, V., *et al.*, eds. 2006. *Large sample basin experiments for hydrological model parameterization: results of the model parameter experiment – MOPEX*. Wallingford, UK: International Association of Hydrological Sciences. IAHS Publ. 307.

Arheimer, B. and Lindström, G., 2014. Electricity vs ecosystems – understanding and predicting hydropower impact on Swedish river flow. *Proceedings of the International Association of Hydrological Sciences*, 364, 313–319. doi:10.5194/piahs-364-313-2014

Arheimer, B. and Lindström, L., 2013. Implementing the EU water framework directive in Sweden. Chapter 11.20. *In*: G. Bloeschl, *et al.*, eds. *Runoff predictions in ungauged basins – synthesis across processes, places and scales*. Cambridge, UK: Cambridge University Press, 353–359.

Arheimer, B., *et al.*, 2011. E-HypeWeb: Service for water and climate information – and future hydrological collaboration across Europe? *In*: J. Hřebíček, G. Schimak, and R. Denzer, eds. *Environmental software systems. Frameworks of eEnvironment*, IFIP Advances in Information and Communication Technology. Berlin: Springer, vol. 359, 657–666.

Berghuijs, W.R., et al., 2014. Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales. *Water Resources Research*, 50 (7), 5638–5661. doi:10.1002/2014WR015692

Beven, K.J., 2000. Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences*, 4 (2), 203–213. doi:10.5194/hess-4-203-2000

Blöschl, G., *et al.*, eds, 2013. *Runoff predictions in ungauged basins – synthesis across processes, places and scales*. Cambridge, UK: Cambridge University Press.

Blöschl, G. and Zehe, E., 2005. On hydrological predictability. *Hydrological Processes*, 19 (19), 3923–3929. doi:10.1002/hyp.6075

Carrillo, G., *et al.*, 2011. Catchment classification: hydrological analysis of catchment behavior through process-based modeling along a climate gradient. *Hydrology and Earth System Sciences*, 15, 3411–3430. doi:10.5194/hess-15-3411-2011

Dee, D.P., *et al.*, 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137, 553–597. doi:10.1002/qj.828

Döll, P., Kaspar, F., and Lehner, B., 2003. A global hydrological model for deriving water availability indicators: model tuning and validation. *Journal of Hydrology*, 270 (1–2), 105–134. doi:10.1016/S0022-1694(02)00283-4

Donnelly, C., *et al.*, 2009. An evaluation of multi-basin hydrological modelling for predictions in ungauged basins. In: K.K Yilmaz *et al.*, eds. New approaches to hydrological prediction in data-sparse *regions* Wallingford, UK: International Association of Hydrological Sciences. IAHS Publ. 333, 112–120.

Donnelly, C., *et al.*, 2013b. Regional overview of nutrient load in Europe – challenges when using a large-scale-model approach, E-HYPE. *In*: B. Arheimer *et al.*, eds. *Understanding freshwater quality problems in a changing world*, Wallingford, UK: International Association of Hydrological Sciences. IAHS Publ. 361, 49–58.

Donnelly, C., Rosberg, J., and Isberg, K., 2013a. A validation of river routing networks for catchment modelling from small to large scales. *Hydrology Research*, 44 (5), 917–925. doi:10.2166/nh.2012.341

Eder, G., Sivapalan, M., and Nachtnebel, H.P., 2003. Modelling water balances in an Alpine catchment through exploitation of emergent properties over changing time scales. *Hydrological Processes*, 17, 2125–2149. doi:10.1002/hyp.1325

Falkenmark, M. and Chapman, T., eds., 1989. *Comparative hydrology*. Paris: UNESCO.

FAO (Food and Agriculture Organization of the United Nations), 1995. *The digitized soil map of the world and derived soil properties (version 3.5)*, FAO land and water digital media series. Rome, Italy: FAO.

Gao, H., *et al.*, 2014. Testing the realism of a topography-driven model (FLEX-Topo) in the nested catchments of the Upper Heihe, China. *Hydrology and Earth System Sciences*, 18, 1895–1915. doi:10.5194/hess-18-1895-2014

Gudmundsson, L., *et al.*, 2012. Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe. *Journal of Hydrometeorology*, 13, 604–620. doi:10.1175/JHM-D-11-083.1

Gupta, H.V., *et al.*, 2014. Large-sample hydrology: a need to balance depth with breadth. *Hydrology and Earth System Sciences*, 18, 463–477. doi:10.5194/hess-18-463-2014

Gupta, H.V., *et al.*, 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377, 80–91. doi:10.1016/j.jhydrol.2009.08.003

Hagemann, S., *et al.*, 2009. Improved regional scale processes reflected in projected hydrological changes over large European catchments. *Climate Dynamics*, 32 (6), 767–781. doi:10.1007/s00382-008-0403-9

Hanasaki, N., *et al.*, 2008a. An integrated model for the assessment of global water resources – Part 1: model description and input meteorological forcing. *Hydrology and Earth System Sciences*, 12, 1007–1025. doi:10.5194/hess-12-1007-2008

Hanasaki, N., *et al.*, 2008b. An integrated model for the assessment of global water resources – Part 2: Applications and assessments. *Hydrology and Earth System Sciences*, 12, 1027–1037. doi:10.5194/hess-12-1027-2008

Hanasaki, N., Kanae, S., and Oki, T., 2006. A reservoir operation scheme for global river routing models. *Journal of Hydrology*, 327, 22–41. doi:10.1016/j.jhydrol.2005.11.011

Hrachowitz, M., *et al.*, 2013. A decade of predictions in ungauged basins (PUB) – a review. *Hydrological Sciences Journal*, 58 (6), 1198–1255. doi:10.1080/02626667.2013.803183

Hundecha, Y., Ouarda, T.B.M.J., and Bárdossy, A., 2008. Regional estimation of parameters of a rainfall–runoff model at ungauged watersheds using the "spatial" structures of the parameters within a canonical physiographic-

climatic space. *Water Resources Research*, 44, W01427. doi:10.1029/2006WR005439

Johansson, B., 2002. *Estimation of areal precipitation for hydrological modelling in Sweden*. Thesis (PhD). Göteborg University. Sweden.

Jothityangkoon, C., Sivapalan, M., and Farmer, D., 2001. Process controls of water balance variability in a large semi-arid catchment: downward approach to hydrological model development. *Journal of Hydrology*, 254 (1–4), 174–198. doi:10.1016/S0022-1694(01)00496-6

Kumar, P., 2007. Variability, feedback, and cooperative process dynamics: elements of a unifying hydrologic theory. *Geography Compass*, 1 (6), 1338–1360. doi:10.1111/j.1749-8198.2007.00068.x

Lehner, B., Verdin, K., and Jarvis, A., 2008. New global hydrography derived from spaceborne elevation data. *Eos, Transactions American Geophysical Union*, 89 (10), 93–94. doi:10.1029/2008EO100001

Liang, X., et al., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, 99 (D7), 14415–14428. doi:10.1029/94JD00483

Lindström, G., *et al.*, 2010. Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. *Hydrology Research*, 41 (3–4), 295–319. doi:10.2166/nh.2010.007

Lopez-Gunn, E., Zorrilla, P., and Llamas, R., 2011. The impossible dream? The Upper Guadiana system: aligning changes in ecological systems with changes in social systems. *In*: J. Lundqvist, ed. *On the water front* [online]. Stockholm: SIWI, Vol. 2, 115–126. Available from: http://www.worldwaterweek.org/on the water front 2010 [Accessed 27 January 2012].

Meier, H.E.M., *et al.*, 2014. Ensemble modeling of the Baltic Sea ecosystem to provide scenarios for management. *Ambio*, 43 (1), 37–48. doi:10.1007/s13280-013-0475-6

Merz, R. and Blöschl, G., 2009. A regional analysis of event runoff characteristics with respect to climate and catchment characteristics in Austria. *Water Resources Research*, 44 (1), 1–19.

Montanari, A., *et al.*, 2013. "Panta Rhei – Everything Flows": change in hydrology and society – The IAHS scientific decade 2013–2022. *Hydrological Sciences Journal*, 58 (6), 1256–1275. doi:10.1080/02626667.2013.809088

Nalbantis, I., *et al.*, 2011. Holistic versus monomeric strategies for hydrological modelling of human-modified hydrosystems. *Hydrology and Earth System Sciences*, 15, 743–758.

Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models. Part I. A discussion of principles. *Journal of Hydrology*, 10, 282–290. doi:10.1016/0022-1694(70)90255-6

Panagos, P., 2006. The European soil database. *GEO: Connexion*, 5 (7), 32–33.

Parajka, J., Blöschl, G., and Merz, R., 2007. Regional calibration of catchment models: potential for ungauged catchments. *Water Resources Research*, 43 (6), 1–16. doi:10.1029/2006WR005271

Pearson, K., 1895. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, 240–242.

Pechlivanidis, I.G., McIntyre, N.R., and Wheater, H.S., 2010. Calibration of the semi-distributed PDM rainfall–runoff model in the Upper Lee catchment, UK. *Journal of Hydrology*, 386 (1–4), 198–209. doi:10.1016/j.jhydrol.2010.03.022

Rihani, S., 2002. *Complex systems theory and development practice: understanding non-linear realities*. London: Zed books.

Rudolf, B. and Schneider, U., 2005. Calculation of gridded precipitation for the global land-surface using in-situ gauge observations. *In*: *Proceedings of the 2nd workshop of the international precipitation working group*. Monterey, CA: IPWG, 231–247.

Samaniego, L., Kumar, R., and Attinger, S., 2010. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, 46, W05523. doi:10.1029/2008WR007327

Savenije, H.H.G., 2010. HESS opinions "Topography driven conceptual modelling (FLEX-Topo)". *Hydrology and Earth System Sciences*, 14, 2681–2692. doi:10.5194/hess-14-2681-2010

Sawicz, K., et al., 2011. Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences*, 15 (9), 2895–2911. doi:10.5194/hess-15-2895-2011

Siebert, S., et al., 2005. Development and validation of the global map of irrigation areas. *Hydrology and Earth System Sciences*, 9, 535–547. doi:10.5194/hess-9-535-2005

Sivapalan, M., 2009. The secret of 'doing better hydrological science': change the question! *Hydrological Processes*, 23, 1391–1396. doi:10.1002/hyp.7242

Stahl, K., *et al.*, 2011. Streamflow data from small basins: a challenging test to high resolution regional climate modeling. *Journal of Hydrometeorology*, 12, 900–912. doi:10.1175/2011JHM1356.1

Strömqvist, J., *et al.*, 2009. Using recently developed global data sets for hydrological predictions. *In*: K.K Yilmaz *et al.*, eds. *New approaches to hydrological prediction in data-sparse regions*, Wallingford, UK: International Association of Hydrological Sciences. IAHS Publ. 333, 121–127.

Strömqvist, J., *et al.*, 2012. Water and nutrient predictions in ungauged basins: set-up and evaluation of a model at the national scale. *Hydrological Sciences Journal*, 57, 229–247. doi:10.1080/02626667.2011.637497

Székely, G.J., Rizzo, M.L., and Bakirov, N.K., 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35, 2769–2794. doi:10.1214/009053607000000505

Thielen, J., *et al.*, 2009. The European flood alert system – Part 1: concept and development. *Hydrology and Earth System Sciences*, 13, 125–140. doi:10.5194/hess-13-125-2009

Wagener, T., 2013. A synthesis framework for runoff predictions in ungauged basins. Chapter 2. *In*: G. Bloeschl, *et al.*, eds. *Runoff predictions in ungauged basins – synthesis across processes, places and scales*. Cambridge, UK: Cambridge University Press, 353–359.

Wagener, T., *et al.*, 2007. Catchment classification and hydrologic similarity. *Geography Compass*, 1 (4), 901–931. doi:10.1111/j.1749-8198.2007.00039.x

Widén-Nilsson, E., Halldin, S., and Xu, C.-Y., 2007. Global water-balance modelling with WASMOD-M: parameter estimation and regionalisation. *Journal of Hydrology*, 340 (1–2), 105–118. doi:10.1016/j.jhydrol.2007.04.002

Appendix A

Definition of flow signatures

In all equations below, x_i refers to the daily discharge on day i, and N to the number of days in the discharge time series. A horizontal bar over a variable refers to its arithmetic mean.

MQ: the long-term arithmetic mean discharge

$$\mathrm{MQ} = \bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} \tag{A1}$$

MQ/A: the long-term mean specific discharge

$$\mathrm{MQ}/A = \frac{\mathrm{MQ}}{A} \tag{A2}$$

where A is the catchment area.

Q_{95}/A and Q_{30}/A: the 95th and 30th percentiles of the discharge time series.

The percentile, P, indicates the value below which a given percentage of the data points in the time series falls. In this study it is calculated as follows: Let v be the ascending ordered dataset $v_1 \le v_2 \le \cdots \le v_N$. The rank, j, of the Pth percentile is calculated as $j = \frac{P}{100}(N-1) + 1$; j is split into its integer component, k, and decimal component, d, such that $j = k + d$. The percentile value at k, v_k, is the value at rank k. The Pth percentile value, v_P, is then calculated as a linear interpolation between k and $k + 1$:

$$v_P = \begin{cases} v_1, & \text{for } k = 0 \\ v_N, & \text{for } k = N \\ v_k + d(v_{k+1} - v_k), & \text{for } 0 < k < N \end{cases} \tag{A3}$$

$$Q_{95}/A = \frac{v_{95}}{A} \quad Q_{30}/A = \frac{v_{30}}{A} \tag{A4, A5}$$

CV, the coefficient of variation

$$\mathrm{CV} = \frac{s_x}{\bar{x}} \tag{A6}$$

where s_x is the sample standard deviation:

$$s_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{A7}$$

r, Pearson's product-moment correlation coefficient:

$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{A8}$$

where x and y are paired datasets (e.g. observed and simulated discharge) (Pearson 1895).

Flash, the flashiness of the daily discharge

$$\mathrm{Flash} = 1 - r_{e,f} \tag{A9}$$

$$e = x_1 + x_2 + \cdots + x_{N-1};$$
$$f = x_2 + x_3 + \cdots + x_N$$

where $r_{e,f}$ is the autocorrelation of a time series (i.e. with itself one day earlier).

Appendix B

Evaluation criteria

In all equations below, z refers to the simulated and o to the observed discharge. As above, a horizontal bar over a variable refers to its arithmetic mean.

Nash-Sutcliffe efficiency, NSE:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(z_i - o_i)^2}{\sum_{i=1}^{N}(z_i - \bar{o})^2} \tag{B1}$$

Kling-Gupta efficiency, KGE:

$$\mathrm{KGE} = 1 - \sqrt{(r_{z,o} - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{B2}$$

$$\alpha = \frac{s_z}{s_o}; \qquad \beta = \frac{\bar{z}}{\bar{o}}$$

where r is Pearson's correlation coefficient (above), α is the ratio of the standard deviations and β is the ratio of the arithmetic means.

Relative volume error, RE:

$$\mathrm{RE} = \left(\frac{\sum_{i=1}^{N}(z_i - o_i)}{\sum_{i=1}^{N} o_i}\right) \times 100 \tag{B3}$$